

UNIVERSIDADE FEDERAL DO RIO GRANDE DO NORTE
DCA0133 - APRENDIZAGEM DE MÁQUINA E MINERAÇÃO DE DADOS
3ª LISTA DE EXERCÍCIO e TRABALHOS- 2025.2

1-) Considere o problema de análise de componentes principais (PCA), isto é, determinar em uma distribuição de dados as componentes que tenham associadas a elas a maior variância e representar as mesmas no espaço de dados formado pelos autovetores da matriz de correlação. Neste sentido considere o seguinte problema.

A tabela abaixo apresenta os dados relativos a amostras de solo. Para cada amostra, tem-se as medidas das porcentagens de areia (X1), sedimentos (X2), argila (X3) e a quantidade de material orgânico (X4). Da referida tabela obtenha as estatísticas descritivas de cada variável, isto é, a média, a mediana, o desvio padrão, os valores máximo e mínimo. Sob estas condições :

a-) Obtenha desta tabela a matriz de correlação.

b-) Desta matriz determine os autovalores ordenados do máximo ao mínimo e os autovetores correspondentes.

c-) Apresente as equações das componentes principais, isto é, cada componente é dada por

$$Y_i = \mathbf{e}_i^t X_i = e_{1i} X_1 + e_{2i} X_2 + e_{3i} X_3 + e_{4i} X_4 \quad i = 1, 2, 3, 4, \text{ onde } e_{ji} \text{ é a componente } i \text{ do autovetor } j.$$

d-) Calcule os percentuais de variância para cada componente e ordene a classificação das variáveis segundo este critério.

Tabela: Dados das amostras de solo (Livro – Análise de dados através de métodos de estatística multivariada – Sueli A. Mingoti)

Amostra	Areia (%):X ₁	Sedimentos(%):X ₂	Argila(%):X ₃	Mat. Orgân(%):X ₄
1	79,9	13,9	6,2	3,3
2	78,5	16,3	7,2	2,5
3	68,9	22,6	8,5	3,6
4	62,2	20,2	17,6	2,8
5	69,2	23,7	7,1	0,9
6	67,8	19,8	12,4	3,8
7	61,3	24,9	13,8	2,2
9	71,6	19,2	9,2	3,6
10	83,7	10,5	5,8	4,4
11	67,1	26,5	6,4	1,4
12	59,8	27,9	12,3	3,5
13	66,7	23,2	10,1	2,9

Amostra	Areia (%): X_1	Sedimentos(%): X_2	Argila(%): X_3	Mat. Orgân(%): X_4
14	72,8	14,5	12,7	1,9
15	60,9	28,9	10,2	1,5
16	61,4	29,2	9,4	2,5
17	75,0	16,8	8,2	3,1
18	80,5	11,9	7,6	3,8
19	71,3	18,5	10,2	2,6
20	56,6	28,9	14,5	2,8
21	55,9	32,8	11,3	3,1
22	61,5	28,1	10,4	2,7
23	59,2	28,4	12,4	2,8
24	76,9	16,3	6,8	2,9
25	58,0	27,6	14,4	3,4

2-) Considere os dados apresentados na tabela abaixo. Determine os centroides dos aglomerados "clusters" presentes nos dados, fazendo uso do algoritmo K-means. Para tanto considere os itens (a) - (c) referentes ao processo de inicialização.

Amostra	x_1	x_2	x_3
1	7.82	-4.58	-3.97
2	-6.68	3.16	2.71
3	4.36	2.19	2.09
4	6.72	0.88	2.80
5	-8.64	-3.06	3.50
6	-6.87	0.57	-5.45
7	4.47	-2.62	5.76
8	6.73	-2.01	4.18
9	-7.71	2.34	-6.33
10	-6.91	-0.49	-5.68
11	6.18	-2.81	5.82
12	6.72	-0.93	-4.04
13	-6.25	-0.26	0.56
14	-6.94	-1.22	1.13
15	8.09	0.20	2.25
16	6.81	0.17	-4.15
17	-5.19	4.24	4.04
18	-6.38	-1.74	1.43
19	4.08	1.30	5.33
20	6.27	0.93	-2.78

- a-) Considere que existam três clusters e a inicialização dos centros seja aleatória
b-) Considere que existam três clusters e a inicialização dos centros seja dada por $\mathbf{m}_1=(1,1,1)^t$, $\mathbf{m}_2=(-1,-1,-1)^t$, $\mathbf{m}_3=(1,1,5)^t$.

Compare o resultado obtido com o item (a) e explique a razão da diferenças, incluindo o número de interações para alcançar a convergência.

3-) A propriedade de ordenação topológica do algoritmo SOM pode ser usada para formar uma representação bidimensional abstrata para fins de visualização de um espaço de entrada de alta dimensionalidade. O objetivo é visualizar os dados de dimensão 8 em um espaço de dimensão 2, constituído pela grade de neurônios. Para investigar esta forma de representação, considere uma grade bidimensional de neurônios que é treinada tendo como entrada os dados oriundos de quatro distribuições gaussianas, C_1 , C_2 , C_3 , e C_4 , em um espaço de entrada de dimensionalidade igual a oito, isto é $\mathbf{x} = (x_1, x_2, \dots, x_8)^t$. Todas as nuvens de dados têm variâncias unitária, mas centroídes (vetores média) diferentes, dados por $\mathbf{m}_1 = (0,0,0,0,0,0,0,0)^t$, $\mathbf{m}_2 = (4,0,0,0,0,0,0,0)^t$, $\mathbf{m}_3 = (0,0,0,4,0,0,0,0)^t$, $\mathbf{m}_4 = (0,0,0,0,0,0,0,4)^t$.

a-) Gere as quatro distribuições gaussianas correspondentes a cada centroíde.

b-) Utilizando uma rede SOM formada por uma grade de neurônio 10x10 neurônio obtenha a representação das quatro distribuições em duas dimensões no SOM.

4-) Considere o problema da clusterização dos país com base na tabela de índices de desenvolvimento de países (Fonte ONU- 2002, Livro – Análise de dados através de métodos de estatística multivariada – Sueli A. Mingoti). Para o processo de identificação de aglomerados (“clusters”) considere um método de clusterização hierárquica, o K-means e o SOM. Compare as três soluções.

Países	Expectativa de Vida	Educação	PIB	Estabilidade Política
Reino Unido	0.88	0.99	0.91	1.10
Austrália	0.90	0.99	0.93	1.26
Canadá	0.90	0.98	0.94	1.24
Estados Unidos	0.87	0.98	0.97	1.18
Japão	0.93	0.93	0.90	1.20
França	0.89	0.97	0.85	1.04
Cingapura	0.88	0.87	0.83	1.41
Argentina	0.81	0.92	0.70	0.55
Uruguai	0.82	0.92	0.65	1.05
Cuba	0.75	0.90	0.40	0.07
Colômbia	0.77	0.75	0.69	0.50
Brasil	0.80	0.73	0.82	0.70
Paraguai	0.75	0.70	0.60	0.40
Egito	0.70	0.62	0.60	0.21
Nigéria	0.44	0.58	0.37	-1.36
Senegal	0.47	0.37	0.45	-0.68
Serra Leoa	0.23	0.33	0.27	-1.56
Angola	0.34	0.36	0.51	-1.98
Etiópia	0.31	0.35	0.32	-2.0

Moçambique	0.24	0.37	0.2	-3.0
China	0.80	0.98	0.95	1.09
Média				
Desvio Padrão				

5-) Apresente um estudo sobre mineração de dados (Data Mining).

6-) Apresente um estudo sobre as técnicas de cluster fuzzy. Considere uma ou mais aplicações.

Trabalhos:

i) - Desenvolva um trabalho sobre Aprendizagem por Reforço Profundo (Deep Reinforcement Learning) considerando aplicações da técnica em problemas de livre escolha.

ii) - Desenvolva um trabalho sobre LLM considerando a máquina de aprendizagem transformer. Apresente aplicações em áreas de livre escolha.

A entrega e apresentação da lista e do trabalho correspondem a um processo da terceira avaliação. Portanto a presença é obrigatória.

O trabalho e a lista podem ser feitos de forma individual ou em grupo com dois componentes.

Na apresentação os componentes serão submetidos a questionamentos sobre a solução da lista e o desenvolvimento do trabalho.

Calendário das Atividades do Final do Curso:

09/12/2025 - Apresentação da Lista

11/12/2025 - Apresentação dos Trabalhos